Robust Classification by a Nearest Mean–Median Rule for Generalized Gaussian Pattern Distributions[1]

M. G. Shevlyakova[a], V. E. Klavdiev[b], and G. L. Shevlyakov[c]

[a] *École Polytéchnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[b] *St. Petersburg State Polytechnic University, ul. Polytechnicheskaya 29, 195251 St. Petersburg, Russia*
[c] *Gwangju Institute of Science and Technology, 1 Oryong-dong, Buk-gu, 500-712, Gwangju, South Korea*
e-mail: m.shevlyakova@gmail.com, klavdiev@stu.neva.ru, shev@gist.ac.kr

Abstract—To provide stability of classification, a robust supervised minimum distance classifier based on the minimax (in the Huber sense) estimate of location is designed for the class of generalized Gaussian pattern distributions with a bounded variance. This classifier has the following low-complexity form: with relatively small variances, it is the nearest mean rule (*NMean*), and with relatively large variances, it is the nearest median rule (*NMed*). The proposed classifier exhibits good performance both under heavy- and short-tailed pattern distributions.

DOI: 10.1134/S1054661808020107

1. INTRODUCTION

In statistical pattern recognition, pattern distributions are usually unknown and may vary in a wide range from short- to heavy-tailed forms. To provide stability of the classification performance under uncontrolled departures from the assumed distribution models and to protect against gross outliers in the data, various approaches have been used, including nonlinear programming methods, advanced neural statistical algorithms, and robust estimation procedures [1–3].

Here we consider the simplest scalar case of supervised classifiers, namely, the minimum distance classifier with the conventional decision rule

$$|x - \hat{\theta}_1| \; </> \; |x - \hat{\theta}_2|, \qquad (1)$$

where x is a pattern and $\hat{\theta}_1$ and $\hat{\theta}_2$ are some estimates of location for the patterns obtained from training sets.

Under Gaussian pattern distributions, the use of the sample means in (1) leads to the optimal *NMean*-rule [3]. As the sample mean is an extremely non-robust estimate, we apply robust minimax (in the Huber sense) estimates in (1) in order to enhance the performance of classification in the conditions of uncertainty of an underlying pattern distribution.

The paper is organized as follows. In Section 2, the robust minimax (in the Huber sense) estimate of location for the class of nondegenerate pattern distributions with a bounded variance is written out. In Section 3, a low-complexity threshold mean–median estimate of location is proposed. In Section 4, robust classifiers based on minimax estimates are introduced and studied. In Section 5, the conclusion is given.

2. MINIMAX ESTIMATES OF LOCATION

Huber's minimax approach [1] in robust estimation can be roughly formulated as follows: in a given class of distributions, the least favorable distribution minimizing Fisher information is determined and the maximum likelihood method for this distribution is then applied. This approach provides a guaranteed accuracy of estimation under departures from the assumptions about an underlying distribution.

Let $x_1, \ldots, x_n$ be i.i.d. random variables with common density $f(x - \theta)$ in a convex class F. Then an M-estimate $\hat{\theta}_n$ of a location parameter θ is defined as a zero of $\sum_{i=1}^{n} \psi(x_i - \hat{\theta}_n) = 0$ with a suitable score function $\psi(x)$ belonging to some class Ψ [1]. The minimax approach implies the determination of the least favorable density f^* minimizing Fisher information $I(f)$ over the class F: $f^* = \arg\min_{f \in F} I(f)$ with subsequent designing of the maximum likelihood estimate (MLE) with the score function $\psi^* = -f^{*\prime}/f^*$. Under rather general conditions of regularity imposed on the classes F and Ψ, $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with variance $V(\psi, f)$ satisfying the minimax property and providing the guaranteed estimation accuracy [1]

$$V(\psi^*, f) \le V(\psi^*, f^*) = \sup_{f \in F} \inf_{\psi \in \Psi} V(\psi, f).$$

Within Huber's minimax approach, the choice of a distribution class F determines all the subsequent stages and the character of the corresponding robust

[1] The text was submitted by the authors in English.

Received November 30, 2007

procedure. In turn, the choice of distribution class depends either on the available prior information about data distributions or on the possibilities of getting this information from the data samples.

In practice there often exists prior information about the pattern distribution, for example, about its moments and quantiles. In order to raise the efficiency of robust procedures, it is advantageous to use such information in the minimax settings by introducing the corresponding distribution classes. Here we consider distributions different from ε-contaminated Gaussian. Henceforth, symmetry and unimodality of distributions are assumed.

In the class $F_1 = \{f : f(0) \geq 1/(2a) > 0\}$ of nondegenerate pattern distributions with a bounded density value at the center of symmetry, the least favorable density is the Laplace [4]

$$f_1^*(x) = \frac{1}{2a}\exp\left(-\frac{|x|}{a}\right).$$

Hence, we have the sign score function $\psi_1^*(x) = \mathrm{sgn}(x)/a$ and the sample median $\mathrm{med}_n x$ as the optimal L_1-norm estimate. The parameter a characterizes the dispersion of the central part of a distribution. It is a very wide class, since any unimodal distribution density with a nonzero value at the center of symmetry belongs to it.

In the class $F_2 = \{f : \sigma^2(f) = \int x^2 f(x)dx \leq \bar{\sigma}^2\}$ of pattern distributions with a bounded variance, the Gaussian density is optimal, [4]

$$f_2^*(x) = \frac{1}{\bar{\sigma}\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\bar{\sigma}^2}\right),$$

with the corresponding linear score function $\psi_2^*(x) = x/\bar{\sigma}^2$ and the sample mean $\bar{x}_n$ as the optimal L_2-norm estimate. Being minimax, the sample mean guarantees a reasonably good accuracy of estimation in class F_2 when distribution variances are really small, and it evidently fails with large variances.

Since the optimal solutions in the classes F_1 and F_2 are qualitatively different, it is advantageous to consider the intersection of these classes:

$$F_{12} = \left\{f : f(0) \geq \frac{1}{2a} > 0, \quad \bar{\sigma}^2(f) \leq \bar{\sigma}^2\right\}. \tag{2}$$

The class F_{12} comprises qualitatively different densities, for example, the Gaussian, heavy-tailed ε-contaminated Gaussian, Laplace, Cauchy (with $\bar{\sigma}^2 = \infty$), short-tailed densities close to the uniform, etc. In this case, the least favorable density simultaneously depends on the two parameters a and $\bar{\sigma}^2$ through their ratio $\bar{\sigma}^2/a^2$, having the Gaussian and Laplace densities as the limit cases linked by the Weber–Hermite family of distributions, which can be rather accurately approximated by the generalized Gaussian densities

$$f_p(x; \beta) = \frac{p}{2\beta\Gamma(1/p)}\exp\left(-\frac{|x|^p}{\beta^p}\right) \tag{3}$$

for $p \geq 1$ [4, 5]. In formula (5), β and p are the scale and shape parameters, respectively.

The corresponding minimax estimate of location has the following three branches: (i) with relatively small variances, it is the sample mean; (ii) with relatively large variances, it is the sample median; and (iii) with relatively moderate variances, it is a compromise between them, namely, the L_p-norm estimate with $1 < p < 2$. However, for the class of generalized Gaussian distributions, we have a very similar result.

Theorem 1 (Shevlyakov and Vilchevski, [4]; p. 78). In the parametric subclass of generalized Gaussian distributions (3) of the class F_{12} of nondegenerate distributions with a bounded variance, the minimax estimate of a location parameter θ is given by the L_p-norm estimate of the following form:

$$\theta = \hat{\theta}_{L_p} = \arg\min_{\theta}\sum_{i=1}^{n}|x_i - \theta|^p, \tag{4}$$

with the power p defined by

$$p = \begin{cases} 2, & \bar{\sigma}^2/a^2 \leq 2/\pi, \\ p^*, & 2/\pi < \bar{\sigma}^2/a^2 < 2, \\ 1, & \bar{\sigma}^2/a^2 \geq 2, \end{cases}$$

where p^* satisfies the following equation

$$\frac{p^{*2}\Gamma(3/p^*)}{\Gamma^3(1/p^*)} = \frac{\bar{\sigma}^2}{a^2}. \tag{5}$$

3. A LOW-COMPLEXITY THRESHOLD MEAN-MEDIAN ESTIMATE

A low-complexity approximation of the minimax estimate (4) given by Theorem 1, henceforth called the *threshold mean–median estimate,* has the form

$$\hat{\theta}_{MM} = \begin{cases} \bar{x}_n, & \bar{\sigma}^2/a^2 \leq \lambda, \\ \mathrm{med}_n x, & \bar{\sigma}^2/a^2 > \lambda, \end{cases} \tag{6}$$

where λ is a threshold value such that $2/\pi \leq \lambda \leq 2$. The optimal choice of a threshold value λ can be done by solving the maximin problem

$$\max_{\lambda \in (2/\pi, 2)}\{\min_{p \in (1, 2)}\mathrm{eff}\hat{\theta}_{MM}(p, \lambda)\} \tag{7}$$

Table 1. The minimal efficiency $\mathrm{eff}\hat{\theta}_{MM}(p, \lambda)$ versus the value of λ

λ	$2/\pi$	0.8	0.9	1.0	1.2	1.4	1.6	2.0
$\min\limits_{p} \mathrm{eff}\hat{\theta}_{MM}(p, \lambda)$	$2/\pi$	0.768	0.824	0.868	0.774	0.690	0.617	0.5

for the asymptotic efficiency $\mathrm{eff}\hat{\theta}_{MM}(p, \lambda)$ of the threshold mean–median estimate of location. The precise result giving the optimal value of the threshold can be formulated as follows.

Theorem 2. In the parametric subclass of generalized Gaussian distributions (3) of the class F_{12} of nondegenerate distributions with a bounded variance, the threshold mean–median estimate of the guaranteed efficiency in the sense of criterion (7) is given by

$$\hat{\theta}^{*}_{MM} = \begin{cases} \bar{x}_n, & \bar{\sigma}^2/a^2 \le 1, \\ \mathrm{med}_n x, & \bar{\sigma}^2/a^2 > 1, \end{cases} \tag{8}$$

or, in other words, $\lambda = \lambda^* = 1$. The guaranteed minimum of efficiency is attained at the saddle point $(p^*, \lambda^*) = (1.407, 1)$ and is equal to the efficiencies of the sample mean and the sample median $\mathrm{eff}\hat{\theta}_{MM}(p^*, \lambda^*) = \mathrm{eff}\bar{x}_n = \mathrm{eff}\,\mathrm{med}_n x = 0.868$.

Proof. The efficiency of the threshold mean–median estimate depends both on the power p and the threshold λ as follows:

$$\mathrm{eff}\hat{\theta}_{MM}(p, \lambda) = \begin{cases} \dfrac{\Gamma^2(1/p)}{p^2\Gamma(3/p)\Gamma(2-1/p)}, & \bar{\sigma}^2/a^2 \le \lambda, \\ \dfrac{1}{\Gamma(1/p)\Gamma(2-1/p)}, & \bar{\sigma}^2/a^2 > \lambda, \end{cases} \tag{9}$$

where the upper and lower branches of (9) are the efficiencies of the sample mean and the sample median, respectively. In turn, the power p is defined through the ratio $\bar{\sigma}^2/a^2$ by (5). Hence, the solution of the inner minimization problem in (7) can be obtained by the direct comparison of the efficiencies of the sample mean and the sample median given the value of λ. These results are displayed in Table 1. From Table 1 it also can be seen that the maximum of the minimal efficiency is attained at $\lambda = 1$. Q. E. D.

From Theorem 2 it follows that the maximum loss of efficiency of the threshold mean–median estimate as compared to the precise estimate (4) of Theorem 1 cannot exceed 13%. However, the low-complexity structure of estimate (8) entirely compensates this loss in its efficiency.

Finally, we deal with the following estimates: the sample mean $\bar{x}_n$, the sample median $\mathrm{med}_n x$, the minimax estimate $\hat{\theta}_{L_p}$ (4), the mean–median estimate $\hat{\theta}^{*}_{MM}$ (8), and their adaptive versions when the characteristics of class F_{12} are estimated from the sample:

$$\hat{\bar{\sigma}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2,$$

$$\hat{a} = (2\hat{\mathrm{f}}(0))^{-1} = \frac{(n+1)(x_{(k+2)} - x_{(k)})}{4},$$

where $x_{(k)}$ is the kth order statistic ($n = 2k + 1$) [4, 5].

The performance of these estimates is studied both in asymptotics and on finite samples of $n = 27$ (this sample size originates from the practical problem of glaucoma diagnostics, for the solution of which our algorithms among others have been used) for the Gaussian, Laplace, Cauchy, uniform, Simpson, and generalized Gaussian distributions. The obtained results confirm the results of the similar studies represented in [4]: on the one hand, the minimax and mean–median estimates exhibit high robustness close to the sample median $\mathrm{med}_n x$ on the heavy-tailed distributions, e.g., the Laplace and Cauchy; on the other hand, they are close to the sample mean $\bar{x}_n$ for short-tailed distributions like the Gaussian, Simpson, and uniform.

4. ROBUST CLASSIFIERS OF THE NEAREST MEAN TYPE: PERFORMANCE EVALUATION

In our study, we consider the following classifiers of the nearest mean type (1) based on: (i) the *NMean*-rule, (ii) the *NMed*-rule, and (iii) the *NMeanMed*-rule, the last having the adaptive version of the threshold mean–median estimate (8).

For binary classification, a Monte Carlo experiment is performed on samples $n = 27$ for the Gaussian, Laplace, uniform, and Simpson pattern distributions with unit variance, as well as for the extremely heavy-tailed Cauchy distribution. The classes differ in location: $|\theta_1 - \theta_2| = 3$.

To characterize the quality of classification, the power of classification P_D is computed for each classifier (in this case, the false alarm probability $P_F = 1 - P_D$) and the number of cycles in Monte Carlo modeling is taken equal to 10000; the results of modeling are exhibited in Table 2.

From Table 2 it can be seen that the *NMeanMed*-rule is a compromise between the *NMean*- and *NMed*-rules,

Table 2. Performance of classifiers

	$P_D(NMean)$	$P_D(NMed)$	$P_D(NMM)$
Gaussian	0.93	0.92	0.93
Cauchy	0.71	0.81	0.80
Laplace	0.90	0.94	0.93
Simpson	0.92	0.92	0.92
Uniform	0.93	0.90	0.93

NMM = NMeanMed.

providing good performance both under short- and heavy-tailed pattern distributions. However, it is much closer in performance to the *NMed*-rule.

For short-tail distributions, the performance of all the classifiers is approximately the same; it becomes different under the heavy-tailed distributions, especially for the Cauchy. In the latter case, the *NMean*-rule has the extremely poor performance. Also we may conclude that the low-complexity robust *NMeanMed*-rule performs quite well on the chosen set of pattern distributions.

5. CONCLUSIONS

A low-complexity robust analogue of the nearest mean classifier based on the new minimax (in the Huber sense) estimate of location has been introduced: as the particular cases, it comprises the nearest mean and the nearest median rules. The proposed classifier demonstrates good performance on a wide set of pattern distributions. Finally, we note that the proposed robust classification rules can be extended to the multivariate case on the basis of a coordinate-wise approach.

REFERENCES

1. P. J. Huber, *Robust Statistics* (Wiley, New York, 1981).
2. A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," IEEE Trans. on Pattern Analysis and Machine Intelligence **22**, 4–37 (2000).
3. R. P. W. Duin, "A Note on Comparing Classifiers," Pattern Recognition Lett. **17**, 529–536 (1996).
4. G. L. Shevlyakov and N. O. Vilchevski, *Robustness in Data Analysis: Criteria and Methods* (Utrecht, VSP, 2002).
5. G. L. Shevlyakov and K. S. Kim, "Robust Minimax Detection of a Weak Signal in Noise with a Bounded Variance and Density Value at the Center of Symmetry," IEEE Trans. on Information Theory **52** (3), 1206–1211 (2006).

Maya Shevlyakova received an MS in Applied Mathematics from St. Petersburg State Polytechnic University, St. Petersburg, Russia in 2006. Her work was devoted to statistical analysis of medical data. At present, she is a M.S. student in applied statistics at École Polytéchnique Fédérale de Lausanne, Lausanne, Switzerland, working on the application of statistical methods to genetics analysis.

Vladimir Klavdiev received an MS in Fluid Mechanics from the Leningrad Polytechnic Institute, Leningrad, Russia in 1971 and a PhD in Engineering Cybernetics from the CVUT, Prague, Czechoslovakia in 1981.

Since 1981 he has been with the Department of Applied Mathematics at St. Petersburg State Polytechnic University as an Associate Professor. His research interests include statistics, data analysis, information theory, and mathematical logic. He has published more than 40 papers.

Georgy Shevlyakov received an MS in Control and System Theory (summa cum laude) and a PhD in Signal Processing and Information Theory from the Leningrad Polytechnic Institute, Leningrad, Russia in 1973 and 1976, respectively. In 1991, he received a Dr. Sci. in Mathematical and Applied Statistics from the St. Petersburg Technical University, St. Petersburg, Russia.

From 1976 to 1979, he was with the Biometrics Group at the Vavilov Research Institute in Leningrad as a Research Associate. From 1979 to 1986, he was with the Department of Mechanics and Control Processes at the Leningrad Polytechnic Institute as a Senior Researcher working in the field of robust statistics and signal processing. From 1986 to 1992, he was with the Department of Mathematics of the St. Petersburg Technical University as an Associate Professor, and from 1992 as a Professor. He is currently a Visiting IT Professor at the Department of Information and Communications, Gwangju Institute of Science and Technology (GIST), Korea. His research interests include robust and nonparametric statistics, data analysis, and queuing and information theory along with their applications to signal processing. He has published a monograph on robust statistics (2002), a textbook on probability and mathematical statistics (2001), and more than 70 papers. He is a member of the IEEE and Bernoulli societies.